

# PATENT ABSTRACTS OF JAPAN

(11)Publication number : 11-003321

(43)Date of publication of application : 06.01.1999

(51)Int.Cl. G06F 15/16

G06F 15/16

G06F 9/46

(21)Application number : 09-155374

(71)Applicant : HITACHI LTD

(22)Date of filing : 12.06.1997

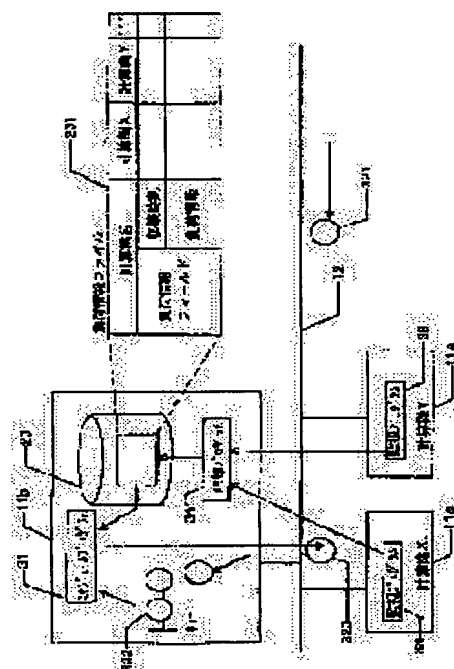
(72)Inventor : HIRATSUKA MASASHI

## (54) PARALLEL COMPUTER SYSTEM

### (57)Abstract:

PROBLEM TO BE SOLVED: To collect load information without raising a load concerning each parallel computer system assigning a job according to the load of a computer for each job.

SOLUTION: Each computer for a job 11a executes a monitoring program 33 collecting load information of the computer for a job 11a in a priority order lower than any other job to transmit to a scheduler computer 11b. In addition the computer 11b executes a recording program 34 recording newest loading information transmitted from a monitoring program 33 with its collecting time by each computer 11a and a schedule program 31 selecting the computer 11a assigning a new job based on load information recorded by the program 34 and its collecting time.



## LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]



## 【特許請求の範囲】

【請求項 1】 割り当てられたジョブを実行する複数のジョブ実行用計算機と、

各々の前記ジョブ実行用計算機にネットワーク経由で接続され、新たなジョブをいずれかの前記ジョブ実行用計算機または自スケジューラ計算機に割り当てるスケジューラ計算機とを具備し、

前記スケジューラ計算機が、各々の前記ジョブ実行用計算機の負荷を表す負荷情報に基づき、新たなジョブを割り当てる計算機を少なくとも 1 つ選択する並列計算機システムにおいて、

各々の前記ジョブ実行用計算機は、

他のどのジョブよりも低い優先順位で自ジョブ実行用計算機の前記負荷情報を収集して前記スケジューラ計算機に送信する負荷情報収集送信手段を備え、

前記スケジューラ計算機は、

前記負荷情報収集送信手段から送信された最新の前記負荷情報をその収集時刻とともに前記ジョブ実行用計算機ごとに記録する負荷情報記録手段と、

新たなジョブが与えられたとき、前記負荷情報記録手段により記録された前記負荷情報及びその収集時刻に基づいて前記新たなジョブを割り当てる計算機を選択する計算機選択手段とを備えることを特徴とする並列計算機システム。

【請求項 2】 前記計算機選択手段は、前記負荷情報記録手段により記録された前記負荷情報及びその収集時刻に基づき、前記収集時刻から現在までに経過した時間を前記ジョブ実行用計算機ごとに求め、

この時間が所定の収集時間間隔を超過しているすべての前記ジョブ実行用計算機を、新たなジョブの割り当て対象から除外することを特徴とする請求項 1 記載の並列計算機システム。

【請求項 3】 前記計算機選択手段は、新たなジョブを割り当てる計算機として、前記時間が前記収集時間間隔を超過していない前記ジョブ実行用計算機のうち、対応する前記負荷情報で表される負荷が最小のジョブ実行用計算機を選択することを特徴とする請求項 2 記載の並列計算機システム。

## 【発明の詳細な説明】

## 【0001】

【発明の属する技術分野】 本発明はジョブを複数の計算機に分散実行させる並列計算機システムに係り、特に、各計算機の負荷情報に応じてジョブの割り当てを行う並列計算機システムに関する。

## 【0002】

【従来の技術】 従来より、複数のジョブ実行用計算機及びこれらのジョブ実行用計算機にジョブを割り当てるスケジューラ計算機をネットワーク経由で相互に接続し、新たに投入されたジョブをスケジューラ計算機が各々のジョブ実行用計算機に分散実行させる並列計算機システ

ムが知られている。こうした並列計算機システムでは、ジョブが投入されたとき、負荷が最も小さいと考えられるジョブ実行用計算機をスケジューラ計算機が選択し、このジョブ実行用計算機に対してジョブの実行を割り当てる、という負荷分散処理を行っている。そして、この負荷分散処理を実現するには、スケジューラ計算機が、各々のジョブ実行用計算機からそのジョブ実行用計算機の負荷を表す負荷情報を収集することが必要である。このため、スケジューラ計算機は、一定時間ごとに、あるいは新たなジョブが投入されたときに、この負荷情報の収集を行っている。

## 【0003】

【発明が解決しようとする課題】 上述した負荷情報の収集を一定時間ごとに行う方法は、ジョブ実行用計算機の負荷が既にかなり高くなっているときでも負荷情報の収集が行われるため、この負荷情報の収集によってジョブ実行用計算機やネットワークの負荷をさらに高めてしまい、これによって、並列計算機システムの性能低下を招くなどの悪影響を及ぼすことがあるという問題点があった。

【0004】 一方、新たなジョブが投入されたときに負荷情報の収集を行う方法は、新たなジョブが投入された時点で、あらゆるジョブ実行用計算機に対して同時に負荷情報の収集が行われるため、ネットワークの負荷を急激に高め、また、それらの収集された負荷情報を解析するスケジューラ計算機に対して短時間に負荷が集中してしまい、これによって、並列計算機システムのスループットが低下することがあるという問題点があった。

【0005】 したがって本発明の目的は、上記方法を用いる従来技術の問題点を解決して、新たなジョブの割り当て対象とするジョブ実行用計算機を選択に必要な負荷情報の収集に際し、ジョブ実行用計算機やネットワークの負荷が急激に高められることのない並列計算機システムを提供することにある。

## 【0006】

【課題を解決するための手段】 上記の目的を達成するため、本発明の並列計算機システムは、割り当てられたジョブを実行する複数のジョブ実行用計算機と、各々の前記ジョブ実行用計算機にネットワーク経由で接続され、新たなジョブをいずれかの前記ジョブ実行用計算機または自スケジューラ計算機に割り当てるスケジューラ計算機とを具備し、前記スケジューラ計算機が、各々の前記ジョブ実行用計算機の負荷を表す負荷情報に基づき、新たなジョブを割り当てる計算機を少なくとも 1 つ選択する並列計算機システムにおいて、各々の前記ジョブ実行用計算機は、他のどのジョブよりも低い優先順位で自ジョブ実行用計算機の前記負荷情報を収集して前記スケジューラ計算機に送信する負荷情報収集送信手段を備え、前記スケジューラ計算機は、前記負荷情報収集送信手段から送信された最新の前記負荷情報をその収集時刻と

もに前記ジョブ実行用計算機ごとに記録する負荷情報記録手段と、新たなジョブが与えられたとき、前記負荷情報記録手段により記録された前記負荷情報及びその収集時刻に基づいて前記新たなジョブを割り当てる計算機を選択する計算機選択手段とを備えるものである。

【0007】

【発明の実施の形態】以下、本発明の並列計算機システムの実施の形態を図面を用いて詳細に説明する。

【0008】図1は本発明の並列計算機システムの一実施形態の構成を示すブロック図、図2は図1のシステムにおける負荷情報の収集について説明する概念図である。両図中、11aは割り当てられたジョブを実行する複数のジョブ実行用計算機、11bは各々のジョブ実行用計算機11aにネットワーク12経由で接続され、新たなジョブをジョブ実行用計算機11aのいずれかに割り当てるスケジューラ計算機、21はスケジューラ計算機11bで実行されるプログラムに応じた処理を行うCPU、22はスケジューラ計算機11bで処理中のデータなどを一時的に格納するメモリ、23はスケジューラ計算機11bでプログラムやデータを格納する外部記憶装置、231は外部記憶装置23に設けられ、後述する記録プログラムにより各々のジョブ実行用計算機11aの負荷を表す負荷情報及びその収集時刻が記録される負荷情報ファイル、24はこれらCPU21、メモリ22、外部記憶装置23の間でプログラムやデータをやりとりするためのバス、31はスケジューラ計算機11bのCPU21で常に実行され、上述した新たなジョブをジョブ実行用計算機11aのいずれかに割り当てる具体的な処理を行う計算機選択手段であるスケジューラプログラム、321は新たに投入されたジョブ、322は割り当て待ちキューに登録された実行割当て待ちジョブ、323は割り当てられた特定のジョブ実行用計算機11aへ送信中の実行待ちジョブ、33は各々のジョブ実行用計算機11aで実行される負荷情報収集送信手段である監視プログラム、34はスケジューラ計算機11bのCPU21で常に実行され、上述の監視プログラム33から送信された最新の負荷情報をその収集時刻とともにジョブ実行用計算機11aごとに負荷情報ファイル231に記録する負荷情報記録手段である記録プログラムである。なお、図1中にはジョブ実行用計算機11aを4台のみ接続した構成を示してあるが、このジョブ実行用計算機11aの接続台数は、必要に応じて所望の台数とすることができる。

【0009】図2において、各々のジョブ実行用計算機11aで実行されている監視プログラム33は、他のどのジョブよりも低い優先順位で、そのジョブ実行用計算機の負荷情報、例えばCPU使用率やメモリ使用率などを収集して前記スケジューラ計算機に送信する。この監視プログラム33による負荷情報の収集及び送信は、一応の目安である後述する所定の収集時間間隔で繰り返し

行われる。すなわち、あるジョブ実行用計算機11aに割り当てられたジョブによるCPU使用率が比較的小かった場合、優先順位の低い監視プログラム33に実行制御が渡される頻度が大きくなるため、監視プログラム33による負荷情報の収集及びそのスケジューラ計算機11bへの送信は、概ね収集時間間隔ごとに行われる。これに対して、上記ジョブによるCPU使用率がある水準を超えて大きくなった場合、優先順位の低い監視プログラム33に実行制御が渡される頻度は小さくなるため、監視プログラム33による負荷情報の収集及びそのスケジューラ計算機11bへの送信は、収集時間間隔より長い間隔で行われる。また、スケジューラ計算機11bで実行されている記録プログラム34は、各々のジョブ実行用計算機11aからネットワーク12経由で送信されてきた上述の負荷情報を、その収集時刻とともにジョブ実行用計算機11aごとに負荷情報ファイル231中に記録する。すなわち、送信された負荷情報の発信元であるジョブ実行用計算機11aが“計算機X”であれば、負荷情報ファイル231中の“計算機X”に対応する負荷情報フィールド内に、送信された負荷情報をその収集時刻、例えばこの負荷情報をネットワーク12から受信したときの時刻とともに記録する。

【0010】図2において、スケジューラ計算機11bで実行されているスケジューラプログラム31は、新たに投入されたジョブ321が割り当て待ちキューに登録されて実行割当て待ちジョブ322となったとき、記録プログラム34により記録された負荷情報ファイル231を参照し、各々のジョブ実行用計算機11aごとに、記録された最新の収集時刻から現在時刻すなわち負荷情報ファイル231を参照した時刻までに経過した時間を求める。そして、この時間が所定の収集時間間隔を超過しているジョブ実行用計算機11aは、すべて実行割当て待ちジョブ322の割り当て対象から除外する。続いて、この時間が所定の収集時間間隔を超過していないジョブ実行用計算機11aについて負荷情報ファイル231を参照し、対応する負荷情報で表される負荷が最小のジョブ実行用計算機11aを、新たな実行割当て待ちジョブ322を割り当てるべきジョブ実行用計算機11aとして選択する。最後に、この実行割当て待ちジョブ322を実行待ちジョブ323として、ネットワーク12経由で選択された特定のジョブ実行用計算機11aへ送信する。

【0011】図3は、図1のシステムにおけるジョブの割り当て対象の選択について説明する概念図である。同図中、横軸は時刻の経過を表し、tは現在時刻、iは一応の目安である所定の収集時間間隔である。ここで、収集時間間隔iの具体的な時間の値は、並列計算機システムの運用方針などに基づいて決めた適当な定数値としてある。また、dは、負荷情報ファイル231に記録された最新の収集時刻から現在時刻tまでに経過した時間であ

る。

【0012】一般に、ジョブ実行用計算機 11a に割り当てられたジョブによる負荷は、ジョブの同時実行数やジョブの処理状態などに応じて常に変化する。そして、上述したように、監視プログラム 33 が負荷情報の収集及び送信を行う間隔は、ジョブ実行用計算機 11a に割り当てられたジョブによる CPU 使用率などの負荷に応じて常に変化する。すなわち、この負荷が比較的小さいとき、負荷情報の収集及び送信の間隔は収集時間間隔  $i$  に概ね一致するが、負荷がある水準を超えて大きくなると、上記間隔は収集時間間隔  $i$  より長くなる。このように、負荷情報の収集及び送信の間隔は常に変化しているが、少なくともこの間隔が収集時間間隔  $i$  より長い場合、該当するジョブ実行用計算機 11a に割り当てられたジョブによる負荷がある水準を超えて大きくなっているものとみなされる。すなわち、図 3 に示すように、負荷情報ファイル 231 に記録された最新の収集時刻から現在時刻  $t$  までに経過した時間  $d$  と収集時間間隔  $i$  との大小関係に対応して、3 種類のケース①②③が生ずるが、このうち、最新の収集時刻から現在時刻  $t$  までに経過した時間  $d$  が収集時間間隔  $i$  より長いケース③は、該当するジョブ実行用計算機 11a に割り当てられたジョブによる負荷がある水準を超えて大きくなっているものとみなされる。したがって、このケース③に該当するジョブ実行用計算機 11a は、すべて新たなジョブの割り当て対象から除外する。次に、残りのケース①②については、該当するジョブ実行用計算機 11a に割り当てられたジョブによる負荷がある水準より小さいことしかわからないので、再び負荷情報ファイル 231 を参照する。そして、記録された負荷情報で表される負荷が最小となっているジョブ実行用計算機 11a を、新たなジョブを割り当てるジョブ実行用計算機 11a として選択する。

【0013】以上のように、本実施形態の並列計算機システムにおいて、各々のジョブ実行用計算機 11a で実行されている監視プログラム 33 は、他のどのジョブよりも低い優先順位でそのジョブ実行用計算機 11a の負荷を表す負荷情報を収集してスケジューラ計算機 11b に送信している。このため、通常の優先順位のジョブの実行に伴ってジョブ実行用計算機 11a の負荷がかなり高くなったときは、より低い優先順位の監視プログラム 33 の実行はほとんど停止し、負荷情報の収集及びスケジューラ計算機 11b への送信が事実上行われなくなるので、既にかなり負荷が高くなっているジョブ実行用計算機 11a やネットワーク 12 の負荷が監視プログラム 33 によってさらに高められることがない。また、スケジューラ計算機 11b で実行されているスケジュールプログラム 31 は、新たなジョブが与えられたとき、既に負荷情報ファイル 231 に記録されている負荷情報のみを参照して、このジョブを割り当てるジョブ実行用計算

機 11a を選択し、各々のジョブ実行用計算機 11a から改めて負荷情報を収集しないため、新たなジョブの登録に際してジョブ実行用計算機 11a やネットワーク 12 の負荷が急激に高められることがない。したがって、新たなジョブの割り当て対象とするジョブ実行用計算機 11a の選択に必要な負荷情報の収集に際し、ジョブ実行用計算機 11a やネットワークの負荷 12 が急激に高められることがなくなり、並列計算機システムの性能低下などの悪影響の発生を防止することができる。

【0014】さらに、新たなジョブの実行を割り当てるジョブ実行用計算機 11a の選択に際し、各ジョブ実行用計算機 11a に対応する最新の収集時刻から現在時刻  $t$  までに経過した時間  $d$  と所定の収集時間間隔  $i$  との比較によって割り当て対象を絞り込んだ後、負荷情報で表される負荷が最小となっているジョブ実行用計算機 11a を選択するので、前回収集された負荷情報による負荷は比較的小さかったが、現時点では負荷がある水準より大きいために時間  $d$  が収集時間間隔  $i$  を超過するジョブ実行用計算機 11a を、割り当て対象から除外することができる。

【0015】なお、上述した実施形態では、スケジューラ計算機 11b が、新たなジョブをジョブ実行用計算機 11a のいずれかに対してのみ割り当てる場合について説明を行ったが、割り当て対象にスケジューラ計算機 11b 自身を含めるようにしてもよい。この場合には、スケジューラ計算機 11b のすべてが上述したケース③に該当したとき、スケジューラ計算機 11b 自身を新たなジョブの割り当て対象として選択するのが適切である。

【0016】

【発明の効果】以上詳しく説明したように、本発明の並列計算機システムによれば、通常の優先順位のジョブの実行に伴ってジョブ実行用計算機の負荷がかなり高くなったときは、より低い優先順位の負荷情報収集送信手段による負荷情報の収集及びスケジューラ計算機への送信が事実上行われなくなるので、既にかなり負荷が高くなっているジョブ実行用計算機やネットワークの負荷が負荷情報収集送信手段によってさらに高められることがない。また、計算機選択手段は、新たなジョブが与えられたとき、負荷情報記録手段によって既に負荷情報ファイルに記録されている負荷情報のみを参照して、このジョブを割り当てるジョブ実行用計算機を選択し、各々のジョブ実行用計算機から改めて負荷情報を収集しないため、新たなジョブの登録に際してジョブ実行用計算機やネットワークの負荷が急激に高められることがない。したがって、新たなジョブの割り当て対象とするジョブ実行用計算機の選択に必要な負荷情報の収集に際し、ジョブ実行用計算機やネットワークの負荷が急激に高められることがなくなり、並列計算機システムの性能低下などの悪影響の発生を防止することができる。

【図面の簡単な説明】

【図 1】本発明の並列計算機システムの一実施形態の構成を示すブロック図である。

【図 2】図 1 のシステムにおける負荷情報の収集について説明する概念図である。

【図 3】図 1 のシステムにおけるジョブの割当て対象の選択について説明する概念図である。

【符号の説明】

11a ジョブ実行用計算機

11b スケジューラ計算機

12 ネットワーク

21 CPU

23 外部記憶装置

231 負荷情報ファイル

31 スケジュールプログラム

322 実行割当て待ちジョブ

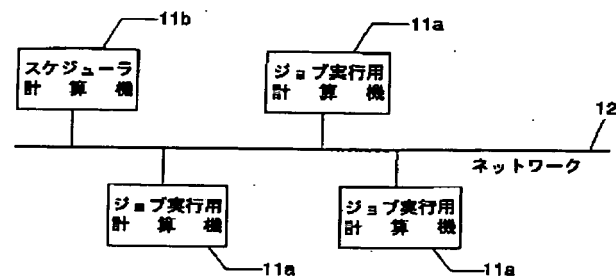
323 実行待ちジョブ

33 監視プログラム

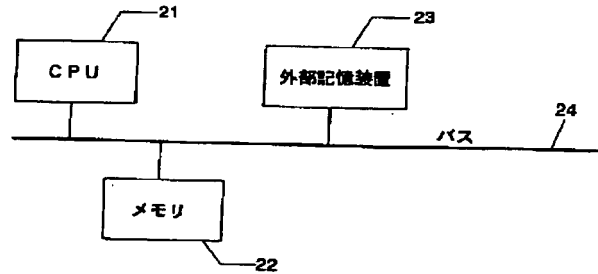
【図 1】

【図 1】

(a)

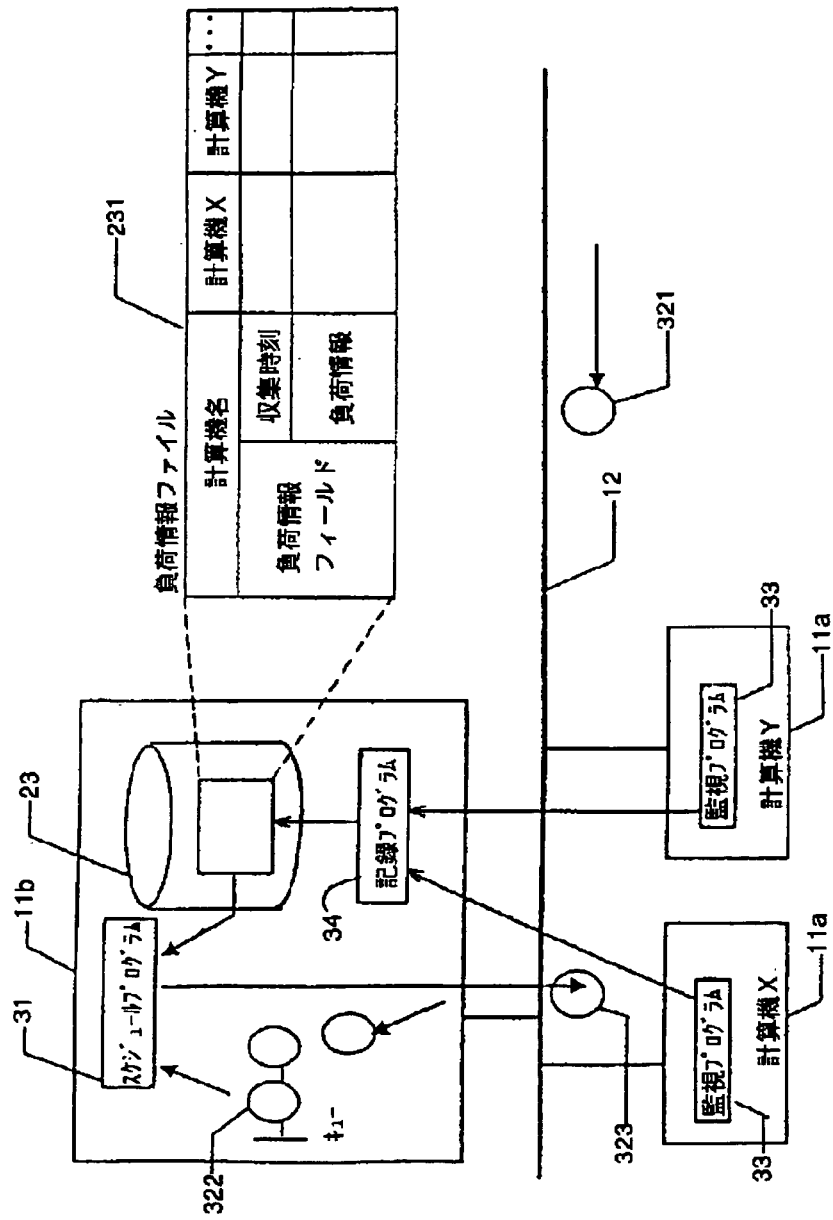


(b)



【図 2】

【図 2】



【図 3】

【図 3】

